A Language for Probabilistic Modeling of Scientific Data

Michael Turmon, Eric Mjolsness, Vladimir Gluzman
Data Understanding Systems Group
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA, USA

Lukman Ramsey
Institute For Neural Computation
University of California San Diego
San Diego, CA, USA

May 25, 2000

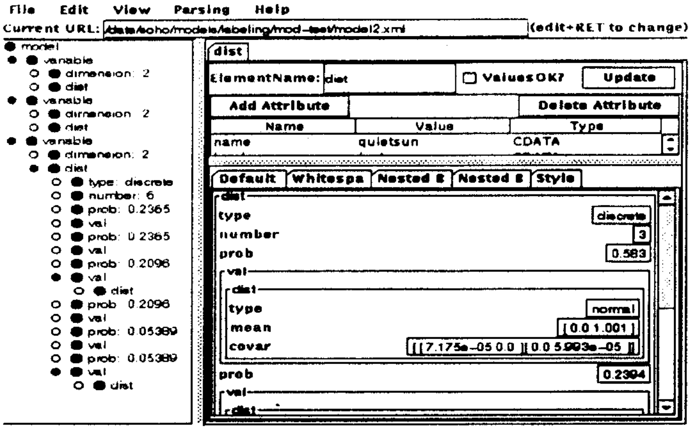
Abstract

We describe a a portable container which allows specification of the probabilistic relations between several variables. The applications motivating this work are mainly scientific inference problems. The textual container should allow users (scientists) to maintain, annotate, edit, and exchange definite models of their data. These specification documents can also be interpreted by inference engines that can automatically infer variables when given the values of other variables. The specification allows hierarchical composition of variables, permitting construction of complex and versatile models from simple ones. Furthermore, special constructs allow easier description of temporal and spatial patterns of dependence between random variables.

1 Analytic foundation

As stated above, our goal is to design an expressive language for capturing stochastic or uncertain relations among a set of variables, particularly in support of the inference tasks associated with pattern recognition in scientific data. At a minimum, we need to be able to describe random vectors via a set of standard distributions, and allow transformations and compositions of these variables. For example, a normal mixture is a composition of a discrete random variable with a group of different normal variables.

One guide to this work has been the recent maturity of Bayes net methods for describing data (Pearl, 1988). This well-known formalism represents random



Current URL: /data/soho/models/labeling/mod-test/model2.xml

```
● model
• ● variable
    ○ ● dimension: 2
    ○ ● dist
• ● variable
    ○ ● dimension: 2
    ○ ● dist
• ● variable
    ○ ● dimension: 2
    • ● dist
        ○ ● type: discrete
        ○ ● number: 6
        ○ ● prob: 0.2365
        ○ ● val
        ○ ● prob: 0.2365
        ○ ● val
        ○ ● prob: 0.2096
        • ● val
            ○ ● dist
        ○ ● prob: 0.2096
        ○ ● val
        ○ ● prob: 0.05389
        ○ ● val
        ○ ● prob: 0.05389
        • ● val
            ○ ● dist
```

ElementName: dist
□ ValuesOK?

Name	Value	Type
name	quietsun	CDATA

```
dist
type      discrete
number    3
prob      0.583
val
  dist
  type    normal
  mean    [0.0 1.001]
  covar   [[7.175e-05 0.0][0.0 5.993e-05]]
prob      0.2394
val
  dist
```

Figure 1: Editing a hierarchical model in JUMBO

variables as the vertices of a directed graph. Lack of an edge between two vertices v_1 and v_2 implies a conditional independence relationship between v_1 and v_2 given the rest of the variables. Our baseline required capability is easy to capture in this notation; furthermore, many pattern recognition tasks such as inference in hidden Markov models or Markov random field models may be captured in this notation.

A parallel line of work is in the area of stochastic grammars (Fu, 1974) and pattern theory (Grenander, 1993). Such grammars model random variables via randomly chosen production rules, and, properly extended, can be a more expressive notation for capturing pattern generation processes (Mjolsness, 1995).

A following influence, allowing us to concisely capture spatio-temporal relations, is to describe a collection of random variables as a *field* over a suitable index set or *domain.* The best example of a domain is the first n integers Z_n, and its cyclic version Z/Z_n. The usual set-theoretic operators are permitted on domains, which allows building index sets for images via cartesian products. The field is then a map from the domain to R^d, and generalizes the concept of random vector. Local spatio-temporal dependences may now be specified by translating a template over the domain.

2 Implementation

There are two basic operations connected with such a stochastic model: sampling, and finding the probability (or probability density) of a given set of variables. This means the setup is amenable to an object-oriented design in which variables are the objects supporting these two methods. Access to the model is given by specifying its URI and giving the names of the variables to be sampled, or the variable names and values if the probability is to be computed. As the "back-end" to this system, we have written a library supporting these operations, suitable for linking with applications needing to evaluate models.

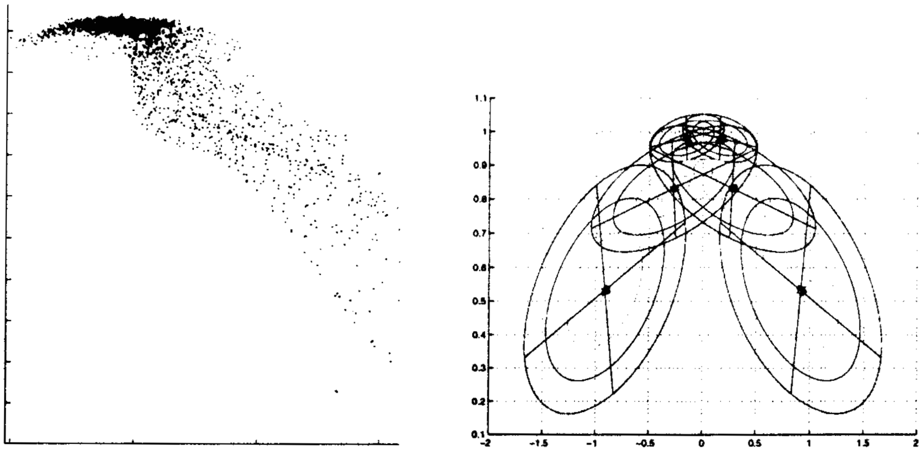


Figure 2: SoHO/MDI observables, and data model

The "front-end" must support browsing and editing. Although many textual representations can capture the above structures, we have chosen XML, the emerging eXtensible Markup Language standard. Using XML allows us to use off-the-shelf parsers to assist the computational engine, as well as standard editors to form the side visible to scientific users. Figure 2 shows one such editing interface, called JUMBO, with a mixture model from a science application highlighted. The model is a set of three variables, each a normal mixture in two dimensions. The variables are encoded, as described above, as a composition of a discrete distribution and several normal distributions.

3 Applications

We describe two applications of this system, the first to solar image analysis and the second to clustering of genetic expression data. In the solar image analysis problem, we observe images of the sun in several modalities, and from these images, we wish to segment the solar disk into region types to explore the links between solar activity and climatic variability.

Many sets of such images are available. The set we describe here consists of magnetograms and photograms of size 1024^2 taken since 1996 from the MDI (Michelson Doppler Imager) instrument aboard the ESA/NASA SoHO satellite. The initial modeling problem is simply to define the statistical distribution of the vector y of two observables for the three region types: sunspot, faculae, and quiet sun. We have done this by fitting a normal mixture to distributions of pixels in scientist-labeled images, obtaining models for $P(y \,|\, \text{spot})$, etc. See figure 2 for a scatterplot of observations y, coded by region type, and a corresponding mixture for the sunspot model. Models for all the classes may be conveniently encoded using the XML scheme we have described. It is precisely such a model which is displayed in figure 1.

We have developed a GUI-driven system called StarTool for this type of image labeling (Turmon *et al.* , 1997). Currently, the system accepts XML probability models which its engine uses to produce labelings of single images, and time-series of labelings of many images. Because of the editing capability described above, it is easy to allow scientists to browse and edit the XML model-files from within the same GUI. This system has the clear advantage that for scientific analysis, models of great complexity can be exchanged between scientists for analysis of images from other instruments, allowing repeatability and objectivity of image-labelings once a modeling consensus forms in the solar physics community. The next problem is to concisely express a Markov random field prior model to account for the spatial dimension, as outlined above.

The second application is in genetic expression determination (Mjolsness (1999)). Here each observable is a measurement of the activity or 'expression' of a certain gene of the nematode worm *C. elegans* in a given experiment. Each experiment in fact determines the activity of many genes, and a collective picture of a gene's behavior emerges as experiments are done over time for several worms. In the data examined here, produced in Stuart Kim's Stanford laboratory, 1244 genes were examined over 68 experiments, and a type of hierarchical clustering was done in R^{68} to group similar genes. For this data, it was found that a *hierarchical* mixture model is more informative about the data than a flat mixture. Such a model is a finite mixture of normal mixtures, and therefore captures information about variability at two scales within the gene sequence. This type of model is again a natural candidate for the hierarchical decompositions supported by our language.

Acknowledgment

This work was carried out by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

References

Fu, K. S. 1974. *Syntactic methods in pattern recognition.* Academic.

Grenander, U. 1993. *General pattern theory: A mathematical study of regular structures.* Oxford.

Mjolsness, E. 1995. *Symbolic neural networks derived from stochastic grammar domain models.* Tech. rept. Dept. of CSE, UC San Diego.

Mjolsness, E. 1999. Clustering methods for the analysis of C. Elegans gene expression array data. *In: Fifth conf. on knowledge discovery and data mining.* Submitted.

Pearl, J. 1988. *Probabilistic reasoning in intelligent systems.* Morgan-Kaufman.

Turmon, M., Mukhtar, S., & Pap, J. 1997. Bayesian inference for identifying solar active regions. *In:* D. Heckerman, H. Mannila, & Pregibon, D. (eds), *Proc. third conf. on knowledge discovery and data mining.* MIT Press.